Disambiguation of English contractions for machine translation of TV subtitles

Disambiguation of English Contractions for Machine Translation of TV Subtitles

Martin Volk and Rico Sennrich
University of Zurich, Institute of Computational Linguistics
Zurich, Switzerland
{volk|sennrich}@cl.uzh.ch

Abstract

This paper presents a disambiguation method for English apostrophe+s contractions. They occur frequently in subtitles and pose special difficulties for Machine Translation. We propose to disambiguate these contractions in a pre-processing step and show that this leads to improved translation quality.

1 Introduction

Ideally, film and TV subtitles are created for each language independently, but for efficiency reasons they are often translated from a source language to one or more target languages. To support the efficient translation we have teamed up with a Scandinavian subtitling company to build Machine Translation (MT) systems. The systems are in extensive practical use today. Because of the established language sequence in the company we have first built translation systems from Swedish to Danish and to Norwegian. After the successful deployment of these two systems, we have worked on other language pairs including English, German and Swedish.

When dealing with English as source language, we have noticed an interesting phenomenon. English subtitles contain a high percentage of contractions. In particular, contractions comprise the short forms of auxiliary and modal verbs: *are* and *were* → *'re*, *shall* and *will* → *'ll*, *had* and *would* → *'d*. By far the most prominent are the apostrophe+s contractions *is*, *was* and *has* → *'s*, which also include contractions of the pronoun *us* (as in *let's do it*) all of which are homographic with the possessive marker. The following table lists the most frequent apostrophe-letter sequences in our corpus of 1 million English subtitles. Note that the last two are dialectal contractions that lead to strange "words" when tokenized at the apostrophe.[1]

172,571	cases of 's	*is,has,us*, possessive
117,869	cases of 't	*not* as in *don't, won't*
53,587	cases of 'm	*am* as in *I'm*
50,219	cases of 're	*are, were*
36,245	cases of 'll	*shall, will*
22,743	cases of 've	*have*
16,576	cases of 'd	*had, should, would*
1,335	cases of 'am	*ma'am*
749	cases of 'all	*y'all*

In this paper we will only be concerned with apostrophe+s contractions because they are the most frequent and the most ambiguous contractions.

Contracted forms are popular in subtitles for several reasons. They are closer to the spoken language of the video, and they are shorter, thus saving precious character space on the screen. Unfortunately, these contractions introduce additional ambiguities into the subtitles which make automatic translation more difficult. We have noticed cases like the following where the English possessive marker *'s* is mistaken by the MT system for the copula verb *is* and therefore mistranslated into Swedish and German.

(1) Oh my gosh, Nicole**'s** dad is the coolest.
SV: Herregud, Nicole **är** pappa är coolast.
DE: Mein Gott, Nicole **ist** Papa ist der coolste.

We have therefore developed a method to disambiguate English apostrophe+s contractions before training the Statistical MT (SMT) system. This paper describes the method and presents the disambiguation results. But first we set the scene

[1] Note that these numbers were computed before lowercasing. The corpus contains also capitalized subtitles like CLEOPATRA'S BEAUTY SALON which we have not counted here.

Bolette Sandford Pedersen, Gunta Nešpore and Inguna Skadiņa (Eds.)
NODALIDA 2011 Conference Proceedings, pp. 238–245

by describing some related disambiguation work and then our MT systems for subtitles.

2 Related Work on Word Sense Disambiguation for MT

Our approach can be seen as a special type of word sense disambiguation (WSD) for MT. Many researchers have worked on this topic before with varying success. For example, (Carpuat and Wu, 2005) reported that they could not find "significant better translations" when using Chinese WSD in Chinese-English MT. But two years later the same authors (Carpuat and Wu, 2007) come to the conclusion that the incorporation of WSD within a typical SMT system "consistently improves translation quality" for Chinese-English. They claim that a disambiguation of phrasal units rather than words leads to these improvements. They report on gains of up to 0.5 BLEU points. These findings are in line with (Chan et al., 2007) who have also shown WSD to be beneficial for Chinese to English translation.

Basic research on WSD for MT is presented in various papers. For example (Specia et al., 2005) work with automatically derived rules for WSD of seven highly ambiguous verbs in English-Portuguese MT. (Apidianaki, 2009) questions the sense inventory which is frequently used in WSD and argues for a semantic analysis based on parallel corpora Greek-English in order to better tailor the sense inventory to MT. (Vickrey et al., 2005) investigate WSD for word translation French-English.

Our work is also similar to other preprocessing suggestions such as (El-Kahlout and Yvon, 2010). They work on the opposite translation direction and prepare the German input text before training and translation to English. When testing various normalization steps, they obtained the biggest improvements on compound splitting.

3 Our MT Systems for TV Subtitles

MT systems for subtitles date back to the work by Popowich et al. (2000) on English to Spanish translation. We have built Machine Translation systems for translating film and TV subtitles from English to Swedish and from Swedish to Danish and Norwegian in a commercial setting. Some of this work has been described earlier by Volk and Harder (2007) and Volk (2008).

Most films are originally in English and receive English or Swedish subtitles in a first manual step. The subtitler uses the English video and audio (sometimes accompanied by an English transcript).

The target language translator subsequently has access to the original English video and audio but also to the source language subtitles and the time codes. In most cases the translator will reuse the time codes and insert the target language subtitle. She can, on occasion, change the time codes if she deems them inappropriate for the target language text.

We have built SMT systems that produce Danish, Norwegian and Swedish draft translations to speed up the translators' work. This project benefited from three favorable conditions:

1. Subtitles are short textual units with little internal complexity.

2. We are dealing with closely related languages. The grammars are similar, however orthography differs considerably, word order differs somewhat and, of course, one language avoids some constructions that the other language prefers.

3. We have access to large numbers of subtitles in multiple languages. The cross-language correspondences can easily be established via the time codes which leads to an alignment on the subtitle level.

There are other aspects of the task that are less favorable. Subtitles are not transcriptions, but written representations of spoken language. As a result the linguistic structure of subtitles is closer to written language than the original (English) speech, and the original spoken content usually has to be condensed by the subtitler.

The task of translating subtitles also differs from most other machine translation applications in that we are dealing with creative language, and thus we are closer to literary translation than technical translation. This is obvious in cases where rhyming song-lyrics or puns are involved, but also when the subtitler applies his linguistic intuitions to achieve a natural and appropriate wording which blends into the video without standing out. Finally, the language of subtitling covers a broad variety of domains from educational

programs on any conceivable topic to exaggerated modern youth language.

We have built SMT systems in order to shorten the development time (compared to a rule-based system) and in order to best exploit the existing translations. We have trained our SMT systems by using standard open source SMT software.

Our corpus consists of TV subtitles from soap operas (like daily hospital series), detective series, animation series, comedies, documentaries, feature films etc. For example, for the Swedish-Danish system we had more than 14,000 subtitle files (= single TV programmes) in each language, corresponding to more than 5 million subtitles (equaling more than 50 million words).

When we compiled our corpus we included only subtitles with matching time codes. If the source and target language time codes differed more than a threshold of 15 TV-frames (0.6 seconds) in either start or end-time, we suspected that they were not good translation equivalents and excluded them from the subtitle corpus. In this way we were able to avoid complicated alignment techniques. Most of the resulting subtitle pairs are high-quality translations thanks to the controlled workflow in the commercial setting. Note that we are not aligning sentences. We work with aligned subtitles which can consist of one or two or three short sentences. Sometimes a subtitle holds only the first part of a sentence which is finished in the following subtitle.

We split our subtitle corpus into training and test set in the usual way. Before the training step we tokenized the subtitles (e.g. separating punctuation symbols from words), converting all uppercase words into lower case, and normalizing punctuation symbols, numbers and hyphenated words.

This resulted, for instance, in BLEU scores of over 50 for the Swedish-Danish system. To better appreciate the translation quality we also calculated exact matches and character-based Levenshtein-5 matches, i.e. subtitles that differ from the human translation by 5 keystrokes or less. We obtained 9% exact matches and 30% Levenstein-5 matches when comparing against a prior human translation. We also ran a number of experiments with post-editors. They got our system output, and we asked them to correct this translation draft into a production-quality subtitle file. When we averaged over 6 post-editors we computed 22% exact matches and 43% Levenstein-5 matches.

In (Volk et al., 2010) we describe the results in more detail. We also present the lessons learned in the process of building the MT systems and integrating them into the workflow of the subtitle company. The systems for Swedish-Danish and Swedish-Norwegian have been in productive use since early 2008 and translate large volumes of subtitles every day. Subsequently we have built a system for English-Swedish translation which went into production in 2010. It was during this latter development that we ran into the problem of contraction ambiguities.

4 The Contraction Ambiguity

In our English-Swedish MT system we observed errors like in example 1 (in the introductory section) which pointed to the problem of translating the various alternatives of apostrophe+s contractions. We started our investigation by manually classifying the occurrences of s-contractions in our subtitle corpus. We identified the following seven variants. An apostrophe+s can be

1. the possessive marker. This case is characterized by the *'s* following a name or a noun and occurring in the beginning of a noun phrase in the article position (i.e. typically in front of a noun or an adjective plus noun; it hardly ever occurs in front of a name). There are rare cases when *'s* follows an indefinite pronoun.

(2) I'm gonna buy Buzzy's store.
You wouldn't say that if we were going after the world's hottest guy.
You even finish each other's sentences.

2. an abbreviation for **the copula 'is'** (or 'was'). This is the most frequent case. It mostly occurs in front of a noun phrase or an adjective (phrase). The distinction between present tense 'is' and the past tense 'was' is only possible in rare sentences with multiple verbs.

(3) Hey, what's your dream, sweetie?
When's the last time you left this place?
-Anything under 60's really slutty.

3. an abbreviation for **the auxiliary 'is'** (or 'was'). This case can be identified by a following verb in present participle form, with perhaps a 'not' or an adverb intervening.

(4) He's trying to find a job.

Michael's thinking about changing his hair.
He's not kidding.
The CIA's still trying to download Dasha...

4. an abbreviation for **the auxiliary 'has'**. This variant occurs in front of a past participle verb form, with perhaps a 'not' or an adverb intervening.

(5) The guy's been really depressed.
For some reason it's lost its magic.

5. an abbreviation for **the auxiliary 'does'** This case is characterized by a question and a verb that is neither a present participle nor a past participle. It is so rare that we do not deal with it.

(6) -What the hell's it look like?

6. an abbreviation for **the pronoun 'us'**. This occurs only after 'let' and can thus trivially be identified.

(7) Let's not rush into this, okay?

7. the plural marker for abbreviations, acronyms and numbers.

(8) I take AP classes and I get all straight A's.
Let me hear those ABC's I taught you.
-With two E's or E-A? -Two E's.
I tolerate them no better on the bench in my 40's.
You know how many number 12's there are on Cold Street?

This plural marker case is difficult to identify. There are many instances of upper-case word plus apostrophe+s or number plus apostrophe+s that do not fall in this category.

(9) Who do you know in the DA's office?
CHP's the last place I belong.
Flight 52's position report is overdue.
Air Traffic Control, flight 52's coming in.

We found that a reliable distinction for acronyms and abbreviations plus apostrophe+s is only possible if they follow a number (*two, three,* ...) or plural indicator like *all* or *many*, or if it occurs at the end of the sentence. For numbers plus apostrophe+s the best indicator is also sentence-final position.

Given these heuristics we identified 131 apostrophe+s occurrences as plural markers after acronyms (upper-case words) and another 546 cases of upper-case word plus apostrophe+s which classify as one of the above alternatives 1-4. For numbers plus apostrophe+s we identified only 15 plural marker cases and 47 others. Since the number of occurrences of both variants is relatively small, we do not handle these cases in our disambiguation.

The most frequent and most difficult ambiguities are between the possessive marker and the copula 'is/was'. The distribution in both cases is similar, and a parse of the sentence would be needed for a precise distinction. All other cases can be disambiguated based on local context and Part-of-Speech (PoS) tags.

5 Disambiguation Method

Since we need PoS information for the disambiguation, we tested different English PoS taggers. It turned out that they do not reliably distinguish the kinds of contractions we are after. For example, the TreeTagger[2] distinguishes between possessive marker (POS), 3rd person singular of the verbs *to be* (VBZ) and *to have* (VHZ),[3] but it never tags apostrophe+s as a pronoun or a plural marker (there isn't even a tag for this). Unfortunately, the TreeTagger does not reliably assign the three tags. For example, it tends to tag apostrophe+s as VBZ (a form of *be*) instead of VHZ when there is an adverb between the contraction and the past participle (see example 10). This is explicable because of the tagger's limited context window. But surprisingly it sometimes also tags the apostrophe+s immediately after the personal pronouns *it* and *he* as possessive markers which can never be correct.

(10) That 's/VBZ always/RB been/VBN a dream of mine.
It/PP 's/POS Sunday morning, ...

Therefore we have developed the following approach:

1. Run a PoS tagger over the subtitles. We used the TreeTagger with the standard English language model.

2. Run a correction script over the tagged subtitles that fixes the apostrophe+s contractions

[2] www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[3] The TreeTagger distinction between VBZ and VHZ is a refinement of the Penn Treebank tag set, which has only one tag for the 3rd person singular form of all verbs.

according to the rules which we sketched in the previous section in the listing of the seven interpretations.

For example, the correction script will turn the apostrophe+s into *is* when it is followed by a word that has been tagged as present participle (*-ing* form). Alternatively it will convert the apostrophe+s into *has* when it is followed by a word that has been tagged as past participle (*-ed* form). Every apostrophe+s that follows *let* is turned into the pronoun *us*. Example 11 shows a subtitle with a copula and a possessive marker before and after disambiguation.

(11) It's always about someone's mother.
It **is** always about someone**'s** mother.

6 Evaluation of the Disambiguation Module

We have tested our disambiguation method on both English-Swedish and English-German SMT systems with similar results. Here we report on our experiments for English-German. For these experiments we obtained 1 million subtitle pairs English-German from our subtitle partner company. The subtitles are of high quality and perfectly aligned on the basis of the time codes. As a first step we tokenized the subtitles by separating punctuation symbols from the words, and by removing the tags for italics and line breaks. This means we do not split or join the subtitles into sentences, instead we regard each subtitle as a translation unit. All subtitles were lower-cased. After tokenization we have 10.3 million tokens in the English subtitles and 8.2 million tokens on the German side.

We checked how many of the English tokens are apostrophe+s tokens. We found a total of 172,571 occurrences of such tokens in 158,328 subtitles. This means around 15% of all subtitles have at least one occurrence of apostrophe+s. These occurrences account for 1.8% of all tokens. After PoS tagging the distribution is as follows. Even if we account for a certain error rate in PoS tagging, it becomes clear that the vast majority of cases are not possessive markers but contractions of auxiliaries and the pronoun *us* (which happens to be tagged as VBZ most of the time).

28,529	tagged as possessive marker (POS)
138,754	tagged as 3rd singular of *be* (VBZ)
5,288	tagged as 3rd sing. of *have* (VHZ)
172,571	Total

In order to appreciate this distribution we compare it to the Penn Treebank. The differences are striking. The Wall Street Journal sections (0-24) of the Penn Treebank have a total of 1.2 million tokens out of which 11,538 are apostrophe+s tokens (0.9% compared to 1.8% in the subtitle corpus). But out of the 49,206 sentences in this treebank, 21% (10,134 sentences) contain such a token. The distribution of their functions is very different from our subtitle corpus. The vast majority (87%) are cases of possessive markers.[4]

10,025	marked as possessive marker (POS)
1,490	marked as 3rd sing. of *be* or *have*
11	marked as personal pronoun (PRP)
12	marked with miscellaneous tags
11,538	Total

From our subtitle corpus we extracted a test set and a development set, around 6500 subtitles each, from across the corpus. The rest (around 990,000 subtitles) was taken as the training set.

Using Moses, we built two SMT systems for English → German translation. The first system was trained on the original subtitles, and the second system was trained on the disambiguated English subtitles and the same German subtitles as before. The disambiguation step changed the English subtitles in the following way:

9493	cases of let's → let us
270	cases of pronoun + 's → has
3644	cases of pronoun + 's → is
1196	cases of other + 's → has
618	cases of other + 's → is

This means, a total of 15,221 PoS tags for apostrophe+s tokens (around 9% of all such tokens) were changed, so that we have the following distribution in our subtitle corpus after correction.

23,540	tagged as possessive marker (POS)
132,788	tagged as 3rd singular of *be* (VBZ)
6,750	tagged as 3rd sing. of *have* (VHZ)
9,493	tagged as personal pronoun (PP)
172,571	Total

In the disambiguation step all occurrences of apostrophe+s that are not tagged as possessive marker (POS) are turned into *is*, *has*, or *us*. Thus our disambiguation substitutes 149,031 oc-

[4] One might wonder whether there are no apostrophe+s occurrences functioning as plural markers in the Penn Treebank. In fact these have been marked with POS, too.

currences (86%) and reduces the apostrophe+s occurrences to the 23,540 possessive markers. Remember that we ignore the apostrophe+s plural markers because they are rare.

Automatic evaluation of both our systems (before and after disambiguation) against the test set of 6510 subtitles resulted in BLEU scores of 28.9. Obviously, BLEU has its limits when tracking small changes in translation. This finding is in line with observations by (Callison-Burch et al., 2006).

Therefore we performed a manual evaluation of the relevant subtitles. Out of the 6510 subtitles in the test set, 1024 subtitles contained apostrophe+s in the original English subtitle. Of these 1024 subtitles our disambiguation module changed 902. This means, in these 902 subtitles an apostrophe+s was turned into *is*, *has* or *us*. But only 224 of these 1024 subtitles have resulted in a different translation than before. We have examined these 224 subtitles in detail and checked whether the translation of the sentences after disambiguation is better than before.

In the following example tables, EN marks the original English subtitle, DE-REF indicates the human-created German reference translation, DE-MT is the output of our MT system before disambiguation. EN-DIS marks the disambiguated English subtitle and DE-DIS-MT the resulting system output.

We found clear cases of improvement as in example table 1. Interestingly, the improvement in this example does not show at the changed copula-apostrophe+s but at the possessive. This is probably due to the fact that the original English subtitles lead to a high translation probability of the apostrophe+s with German *ist* ($prob(ist|'s) = 0.605$), as this is by far the most frequent translation correspondence. This results in the erroneous translation of the apostrophe+s with the German copula *ist*. After disambiguation the probability of apostrophe+s (= possessive marker) with *ist* is much lower (0.319), thus paving the way for the correct German translation with the genitive form of the indefinite pronoun *jemand*.

There are other cases of improvement that are directly related to the disambiguation. Example table 2 shows an improvement for the translation of *'s been* after it has been turned into *has been*. The sentence is still not perfectly translated (mainly because the English word *block* needs to be translated differently in this context), but the translation of the copula verb and the subsequent word order are clearly better.

There are other examples that show worse translations. In particular we find worse translations in connection with *let's* (as in example table 3). We suspect that *let's* is so idiomatic that a split will give too much significance to the pronoun *us* and "disturb" the translation probabilities.

It is also striking that sometimes the disambiguation leads to translations that are different but as good (or bad) as before. Obviously the disambiguation step leads to slight shifts in the translation probabilities that result in changed preferences for one translation over the other. Example 4 is such a case in point with a good idiomatic translation both before and after the disambiguation. Note that the latter translation will receive much higher BLEU scores because of its almost perfect overlap with the reference translation.

Table 5 contains the numerical results of the manual evaluation. Almost 26% of the examined subtitles show an improved translation in relation to apostrophe+s. This stands against about 6% that show a worse translation. So this is a net improvement for 20% of the examined subtitles (which account only for 22% of the subtitles with apostrophe+s in the test set).

Interestingly, we also find translation improvements that are seemingly unrelated to the apostrophe+s in the subtitle since they appear in a different part of the subtitle. We identified 16.5% improvements versus 12% degradations in this class. This adds to the positive overall effect of the disambiguation. The remaining 39% of the subtitles have resulted in translations that are different than before but are judged as being of equal quality (as in example table 4).

These numbers refer only to those 22% of the apostrophe+s-containing subtitles whose MT output had changed after the disambiguation. But the apostrophe+s disambiguation influences also the translation of subtitles without apostrophe+s because of differing word alignments. In order to see how the disambiguation step influences those subtitles, we also manually checked 224 subtitles with changed MT output in this class. There we found no statistically significant difference in translation quality before and after disambiguation.

EN:	it 's always about someone 's mother .
DE-REF:	es hat immer mit der mutter zu tun .
DE-MT:	es geht immer nur um jemand ist mutter .
EN-DIS:	it **is** always about someone 's mother .
DE-DIS-MT:	es geht immer nur um jemandes mutter .

Table 1: Example of improved MT for the possessive marker

EN:	this car 's been on my block for a week .
DE-REF:	seit einer woche steht ein auto in meiner straße.
DE-MT:	das auto ist auf mich block für eine woche .
EN-DIS:	this car **has** been on my block for a week .
DE-DIS-MT:	das auto war auf dem block für eine woche .

Table 2: Example of improved MT for *'s been → has been*

EN:	now , let 's step into the bar .
DE-REF:	treten sie ein .
DE-MT:	also , gehen wir in die bar .
EN-DIS:	now , let **us** step into the bar .
DE-DIS-MT:	lass uns in der bar .

Table 3: Example of worse MT for *let's → let us*

EN:	he 's out of his mind .
DE-REF:	er hat wohl den verstand verloren .
DE-MT:	er ist durchgeknallt .
EN-DIS:	he **is** out of his mind .
DE-DIS-MT:	er hat den verstand verloren .

Table 4: Example of equally good MT for an idiomatic expression

subtitles	percent	human judgement
58	25.9%	better translation related to apostrophe+s
14	6.3%	worse translation related to apostrophe+s
37	16.5%	better translation but not related to apostrophe+s
27	12.1%	worse translation but not related to apostrophe+s
88	39.3%	translation is different, but as good or as bad as before

Table 5: Results of the manual evaluation of 224 subtitles

7 Conclusion

We have shown that film and TV subtitles in general are well suited for MT. But they also have specific properties that make MT more difficult than for other genres. As an example of this, we have investigated apostrophe+s contractions in English that are frequent in subtitles and introduce additional ambiguities.

We have presented a method that disambiguates these contractions based on PoS tags assigned by a general-purpose PoS tagger. We found that this disambiguation has a positive impact on the translation quality of the respective subtitles (although this impact is not visible in the BLEU scores).

On the practical side we plan to investigate whether we can train a PoS tagger to reliably classify the apostrophe+s contractions directly so that we no longer need a separate disambiguation module. One option would be to train a special-purpose PoS tagger on the automatically corrected output of the general-purpose PoS tagger.

In a broader perspective our work reopens the question of whether other disambiguation steps in pre-processing (e.g. for other contraction types) will be similarly beneficial for MT.

Acknowledgments

We would like to thank Nicole Michel for help in assessing the MT output quality.

References

Marianna Apidianaki. 2009. Data-driven semantic analysis for multilingual WSD and lexical selection in translation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 77–85, Athens.

Christopher Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the Thirteenth Conference of the European Chapter of the Association for Computational Linguistics*, Trento.

Marine Carpuat and Dekai Wu. 2005. Word sense disambiguation vs. statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 387–394, Ann Arbor.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 61–72, Prague.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, pages 33–40, Prague.

Ilknur Durgar El-Kahlout and Francois Yvon. 2010. The pay-offs of preprocessing for German-English statistical machine translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation*, Paris.

Fred Popowich, Paul McFetridge, Davide Turcato, and Janine Toole. 2000. Machine translation of closed captions. *Machine Translation*, 15:311–341.

Lucia Specia, Maria das Graças V. Nunes, and Mark Stevenson. 2005. Exploiting rules for word sense disambiguation in machine translation. *Procesamiento del Lenguaje Natural*, (35):171–178.

David Vickrey, Luke Biewald, Marc Teyssier, and Daphne Koller. 2005. Word-sense disambiguation for machine translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 771–778, Vancouver.

Martin Volk and Søren Harder. 2007. Evaluating MT with translations or translators. What is the difference? In *Proceedings of Machine Translation Summit XI*, Copenhagen.

Martin Volk, Rico Sennrich, Christian Hardmeier, and Frida Tidström. 2010. Machine translation of TV subtitles for large scale production. In *Proceedings of the Second Joint EM+/CNGL Workshop on "Bringing MT to the User: Research on Integrating MT in the Translation Industry"*, pages 53–62, Denver.

Martin Volk. 2008. The automatic translation of film subtitles. A machine translation success story? In Joakim Nivre, Mats Dahllöf, and Beáta Megyesi, editors, *Resourceful Language Technology: Festschrift in Honor of Anna Sågvall Hein*, volume 7 of *Studia Linguistica Upsaliensia*, pages 202–214. Uppsala University, Humanistisk-samhällsvetenskapliga vetenskapsområdet, Faculty of Languages.

ISSN 1736-6305 Vol. 11
http://hdl.handle.net/10062/16955